Journal of King Saud University – Computer and Information Sciences (2017) **29**, 462–472

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Intrusion detection model using fusion of chi-square feature selection and multi class SVM

Sumaiya Thaseen Ikram [a,*], **Aswani Kumar Cherukuri** [b]

[a] *School of Computing Science and Engineering, VIT University, Chennai, Tamil Nadu, India*
[b] *School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, India*

Received 7 July 2015; revised 4 October 2015; accepted 3 December 2015
Available online 31 March 2016

KEYWORDS

Chi square feature selection;
Cross validation;
Intrusion detection;
Radial basis kernel;
Support vector machine;
Variance

Abstract Intrusion detection is a promising area of research in the domain of security with the rapid development of internet in everyday life. Many intrusion detection systems (IDS) employ a sole classifier algorithm for classifying network traffic as normal or abnormal. Due to the large amount of data, these sole classifier models fail to achieve a high attack detection rate with reduced false alarm rate. However by applying dimensionality reduction, data can be efficiently reduced to an optimal set of attributes without loss of information and then classified accurately using a multi class modeling technique for identifying the different network attacks. In this paper, we propose an intrusion detection model using chi-square feature selection and multi class support vector machine (SVM). A parameter tuning technique is adopted for optimization of Radial Basis Function kernel parameter namely gamma represented by 'ϒ' and over fitting constant 'C'. These are the two important parameters required for the SVM model. The main idea behind this model is to construct a multi class SVM which has not been adopted for IDS so far to decrease the training and testing time and increase the individual classification accuracy of the network attacks. The investigational results on NSL-KDD dataset which is an enhanced version of KDDCup 1999 dataset shows that our proposed approach results in a better detection rate and reduced false alarm rate. An experimentation on the computational time required for training and testing is also carried out for usage in time critical applications.

1. Introduction

Intrusion detection identifies computer attacks by observing various records processed on the network. Intrusion detection models are classified into two variants, misuse detection and anomaly detection systems. Misuse detection can discover intrusions based on a known pattern also known as signatures (Ilgun et al., 1995). Anomaly detection can identify the malicious activities by observing the deviation from normal network traffic pattern (Sumaiya Thaseen and Aswani

* Corresponding author.
E-mail addresses: sumaiyathaseen@gmail.com (I. Sumaiya Thaseen), aswanis@gmail.com (C. Aswani Kumar).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2015.12.004

Kumar, 2014; Amiri et al., 2011). Hence anomaly detection can identify new anomalies. The difficulty with the current developmental techniques is the high false positive rate and low false negative rate (Sarasamma et al., 2005).

Most of the data mining and bio-informatics applications require processing of large data. A large amount of resources have been utilized in Intrusion Detection Systems (IDS) and several machine learning techniques like decision tree (Lee et al., 2008), genetic algorithm (Shafi and Abbass, 2009), Support vector machines (Khan et al., 2007), Artificial Neural Network (Wang et al., 2010) and hybrid intelligent system (Peddabachigari et al., 2007) are explored to build an IDS. However none of the techniques are able to identify all intrusion attempts and result in a higher detection rate and lower false alarm rate (Panda et al., 2011). Hence there is a need to integrate feature selection and classifier techniques to achieve a better performance.

A model can be learned using supervised or unsupervised learning. Supervised learning requires that the target variable is well known and a sufficient number of values are provided. In unsupervised learning either the target variable is unknown or has been observed only for small number of data.

Support vector machine (SVM) is one of the supervised learning models that has a higher classification efficiency in comparison to other classifier models but due to the higher training time for large data sets, the usage is limited. Hence many feature selection techniques are integrated with SVM to obtain reduced dimensional data. This results in less training time for the classifier. Feature selection is used to select an optimal subset of features for model construction. The feature selection process calculates the score of each probable feature based on a specific feature selection technique and then identifies the best '*k*' features. This procedure is carried out by generating a ranked list of features and different selection criteria can be considered to select a subset of features.

One of the common statistical techniques is the chi-squared that estimates discrepancy from the expected distribution if the feature incidence is not dependent on the class value.

In this paper we put forward an intrusion detection model integrating chi-square feature selection and multi class support vector machine for high accuracy and low false positive rate. The kernel parameter is optimized by obtaining the variance for each attribute feature and determining the highest attribute variance. As the result if kernel is inversely dependent to the variance, a high variance will result in a better kernel parameter. We call this technique as the variance tuning technique.

Many intrusion detection models have been developed with feature selection and classification techniques. The uniqueness of the proposed model over existing intrusion detection approaches is that the optimization of SVM parameters is performed using a variance tuning technique. The variance tuning technique results in a better accuracy in the SVM classifier with minimum time complexity which is detailed in Section 5.1. The average accuracy achieved for all the attacks and normal traffic is more than 95% whereas only the U2R attack accuracy is less as the number of samples involved in training the model is less.

The rest of the paper is structured as follows. The review of various machine learning techniques employed for intrusion detection and the importance of SVM technique for classification along with other feature selection techniques integrated with SVM are introduced in Section 2. The background of various techniques used in the model is detailed in Section 3. The proposed methodology is discussed in Section 4. The experiments and results of the model are reported in Section 5. Section 6 contains the conclusion.

2. Related work

Many hybrid intrusion detection models have been developed to overcome the restrictions of anomaly and misuse detection models. We will analyze the literature of traditional intrusion detection techniques, intrusion models using data mining techniques, intrusion models using single SVM classifiers and integrated intrusion models using SVM and feature selection techniques.

The various techniques used by IDS are statistic (Lazarevic et al., 2003), hidden markov model (Ye and Borror, 2004), artificial neural network (Fisch et al., 2010; Novikov , 2006), fuzzy logic (Saniee Abadeh et al., 2007; Toosi and Kahani, 2007) and rule learning (Xuren et al., 2006). Research in the recent years indicate that SVM can be used for building an intrusion detection model effectively. Fisch et al. (2010) and Mukkamala (2005) have observed the performance of support vector machine, multi variate adaptive regression splines(MARS) and artificial neural network(ANN). It is preferable to build an assembly of classifiers like ANN, MARS and SVM to improve the detection accuracy. Zhang and Shen (2005) used SVM for building an intrusion detection. The system employed text processing methods based on occurrence of system call implemented by the program. Horng et al. (2010) developed a network intrusion detection model using SVM and integrated with BRICH hierarchical clustering for preprocessing. The grouping process reduced the data set thereby decreasing the training time and hence SVM classifiers resulted in higher performance. Ilgun et al. (1995) employed rule based techniques to design and develop IDS, where the expert knowledge is considered as a rule set. Lee et al., (1999) used the data mining technique to create association rules instead of human experts as an analytical model. The drawback of such methods is a large number of association rules are defined thus increasing the complexity of the model.

Due to the large dimensionality of network data, many intrusion models were developed with feature selection considered as a step of preprocessing. Mukkamala (2005) deployed a feature selection technique during preprocessing. At every instance, one input feature is disassociated from the dataset while the residual data set is employed for training and testing. The features are graded based on a set of rules pertaining to the classifiers performance before and after feature selection. Chebrolu et al. (2005) categorized primary features in constructing an IDS that is very crucial for real world detection. Markov model and decision tree has been used in the feature selection process. Bayesian network combined with regression trees were used to build the intrusion detection model. Sung and Mukkamala (2003) eliminated one feature at every time instance to conduct experiments on SVM integrated with neural network. The authors used only 34 significant features rather than all 41 feature sets and obtained a significant performance change in the intrusion detection. Zaman (2009) developed a feature selection technique to construct a lightweight IDS. The proposed approach employed a fuzzy enhanced support vector